Using Regression to Combine Data Sources for Semantic Music Discovery

Brian Tomasik, Joon Hee Kim, Margaret Ladlow, Malcolm Augat, Derek Tingle, Richard Wicentowski, Douglas Turnbull

SWARTHMORE

Introduction

Goal

Create a semantic music discovery engine

- given a text query, return a ranked list of relevant songs

Problem

Need to "annotate" music with tags:

$\widehat{y}_{s,t}$ = (estimated) affinity score between song *s* and tag *t*

Approach

1. Collect information from multiple input data sources (each represented with a score $x_{s,t}$)
2. Combine sources using regression

Research Question

Can we share information across tags using Bayesian hierarchical regression models to improve semantic music annotation?

Data Sources

Web Documents (WD)

Idea

Text-mine tags from relevant web documents

Approach

1. Collect top 10 web pages for each song from Google using *"song name" "artist name"* query
2. Calculate score:

$$x^{WD}_{s,t} = \sum_{d \in D_s} \frac{n_{t,d}}{N_{t,d}}$$

D_s - the set of documents for song s

$n_{t,d}$ ~ # of times tag t appears in document d

$N_{t,d}$ ~ # number of times it could have appeared.

Content-based Autotagging (CB)

Idea

Learn a joint probabilistic model of audio features and tags

Approach

1. Learn a Gaussian mixture model (GMM) distribution over an MFCC feature space for each tag
2. Estimate the posterior probability of tag *t* given bag-of-MFCC vectors (X_s) for song *s*

$$x^{CB}_{s,t} \approx P(t|\mathcal{X}_s) = \frac{P(\mathcal{X}_s|t)P(t)}{P(\mathcal{X}_s)}$$

Top performing approach in 2008 MIREX Audio Tag Classification task [1]

Collaborative Filtering (CF)

Idea

Copy tags from annotated songs to an unannotated song based on artist similarity

Approach

1. Estimate the similarity between two artists based on co-occurrence within the music preference lists of 400,000 Last.fm users.
2. For artist *a* of song *s*, find the set of all songs (*S*) from the closest *k = 32* artists to *a*.
3. Calculate the fraction of songs in *S* that are labeled with tag *t*

$$x^{CF}_{s,t} = \frac{\sum_{i \in S} y_{i,t}}{|S|}$$

See our ISMIR 2009 Paper on Tag Propagation for details [2]

Combination Methods

Goal

Improve estimated affinity scores ($\widehat{y}_{s,t}$) by combining scores from the data sources ($x^{WD}_{s,t}, x^{CB}_{s,t}, x^{CF}_{s,t}$)

Preprocessing Scores

For each tag *t* and data source *i* in {WD, CB, CF}, we

1. Transform scores so they are roughly normally distributed (e.g., log or power transform)
2. Standardize (mean = 0, variance = 1)

Fixed Combiners

- Simple functions of input scores
- E.g., max, min, product, sum, median

e.g. $\rightarrow$ $\widehat{y}_{s,t} = \max(x^{WD}_{s,t}, x^{CB}_{s,t}, x^{CF}_{s,t})$

Drawback: each data source receives "equal" weight

Learned Combiners

- Learn a parametric model using human-labeled training data
- For example, learn *"betas"* for a linear discriminant function

$$\widehat{y}_{st} = \sum_{i \in \{WD,CB,CF\}} \beta^i_t x^i_{st}$$

Linear and Logistic Regression

- Generalized linear models
- Learn beta parameters using maximum likelihood estimation
- Beta parameter for each tag is learned independently from one another

Linear

$$\begin{aligned} y_{s,t} &= \beta_t x_{s,t} + \epsilon_{s,t} \\ \epsilon_{s,t} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_t) \end{aligned}$$

Logistic

$$\begin{aligned} y_{s,t} &= \frac{1}{1 + exp(-z_{s,t})} \\ z_{s,t} &= \beta_t x_{s,t} + \epsilon_{s,t} \\ \epsilon_{s,t} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_t) \end{aligned}$$

Hierarchical Bayesian Models [3]

- Assume beta's share common structure across the vocabulary of tags

$$\begin{aligned} \beta_t &= \overline{\beta} + v_t \\ v_t &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

- For example, three tags with independent beta coefficients equal to 0.1, 0.2, and 0.3 then $\overline{\beta}$ = 0.2 with σ = 0.1

Mixture Hierarchical Model

- Instead of assuming that v_t is normally distributed, assume that v_t comes from a *mixture* of normal distributions
- Intuition: Suppose β_t for "genre tags" clusters around 0.25 and for "acoustic tags" clusters around 0.05 then if $\overline{\beta}$ = 0.20, then v_t might have two peaks at 0.05 and -0.15.

Setup

10,870 songs

2 tag vocabularies

- 71 genre and subgenre tags (e.g., *"rock", "delta blues", "trance", "piano concerto"*)
- 151 acoustic tags from Pandora's Music Genome Project (e.g., *"acoustic instrumentation", "vocal harmonies", "major key tonality"*)

3 data sources (WD, CF, CB) plus popularity (P) value based on Last.fm scrobble count

5-fold cross validation with an artist filter

- Training set split 3-to-1 to first train CB system, and then regression model

Rank order test set song once for each tag -> calculate standard IR evaluation metric s

Results

Regression Model - Independent Linear Regression

	71 Genre Tags				151 Acoustic Tags			
	AUC	MAP	R-Prec	10-Prec	AUC	MAP	R-Prec	10-Prec
Random	0.502±0.003	0.09±0.01	0.08±0.01	0.08±0.02	0.508±0.003	0.032±0.003	0.030±0.003	0.03±0.00
WD	0.666±0.010	0.25±0.02	0.29±0.02	0.47±0.03	0.616±0.006	0.135±0.007	0.181±0.008	0.29±0.02
CF	0.732±0.010	0.45±0.02	0.45±0.02	0.72±0.04	0.641±0.008	0.154±0.010	0.213±0.011	0.25±0.02
CB	0.781±0.014	0.23±0.02	0.25±0.02	0.38±0.03	0.836±0.008	0.141±0.007	0.161±0.008	0.19±0.01
All3	0.871±0.007	**0.52±0.02**	0.50±0.02	**0.74±0.04**	**0.888±0.006**	0.276±0.010	0.298±0.010	0.42±0.02
All3&P	**0.876±0.007**	**0.52±0.02**	**0.51±0.02**	**0.74±0.04**	0.887±0.006	**0.277±0.010**	**0.299±0.010**	**0.42±0.02**

Combination Method - All3&P

	71 Genre Tags				151 Acoustic Tags			
	AUC	MAP	R-Prec	10-Prec	AUC	MAP	R-Prec	10-Prec
Min	0.658±0.015	0.27±0.02	0.27±0.02	0.60±0.04	0.654±0.009	0.121±0.006	0.161±0.008	0.26±0.01
Product	0.826±0.009	0.42±0.03	0.41±0.02	0.67±0.04	0.814±0.006	0.197±0.008	0.232±0.009	0.32±0.01
Median	0.826±0.009	0.43±0.02	0.43±0.02	0.68±0.04	0.820±0.006	0.219±0.009	0.261±0.009	0.35±0.02
Sum	0.851±0.007	0.44±0.03	0.44±0.02	0.69±0.04	0.847±0.006	0.220±0.009	0.252±0.009	0.34±0.01
Max	0.856±0.007	0.46±0.02	0.48±0.02	0.59±0.03	0.859±0.006	0.239±0.009	0.274±0.009	0.34±0.01
Ind Log	0.866±0.006	0.51±0.03	0.50±0.02	0.72±0.04	0.875±0.005	0.266±0.010	0.293±0.010	0.40±0.02
Hier Log	0.872±0.006	0.51±0.03	0.50±0.02	0.73±0.04	0.883±0.006	0.272±0.010	0.296±0.010	0.40±0.02
Hier Mix	**0.876±0.007**	**0.52±0.02**	**0.51±0.02**	**0.74±0.04**	**0.887±0.006**	**0.277±0.010**	**0.299±0.010**	**0.42±0.02**
Hier Lin	**0.876±0.007**	**0.52±0.02**	**0.51±0.02**	**0.74±0.04**	**0.887±0.006**	**0.277±0.010**	**0.299±0.010**	**0.42±0.02**
Ind Lin	**0.876±0.007**	**0.52±0.02**	**0.51±0.02**	**0.74±0.04**	**0.887±0.006**	**0.277±0.010**	**0.299±0.010**	**0.42±0.02**

Conclusions

1) CB is best for AUC metric, CF is best for Precision metrics

- CF and WD produce sparse annotations -> random ranking after first couple of songs

2) CB better relative performance to CF on acoustic tags

- Suggests that genre labels may be more socially-oriented that acoustic tags

3) Popularity information was not too helpful

- Suggests Pandora tags are not biased by popularity

4) **Three data sources are better than one** or two data sources alone

- Data sources are largely uncorrelated for most tags (i.e., corr. coef. < 0.1)
- Beta coefficients are significantly non-zero and positive

5) Trained regression models outperform fixed combiner functions

- But max and sum are not too bad

6) **Independent Linear Regression works** as well as more complex hierarchical models

- Easy to implement, fast to compute, easy to parallelize
- Hierarchical models require additional exploration for cases where there are only a few dozen labeled songs for a tag

[1] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. *Semantic Annotation and Retrieval of Music and Sound Effects.* IEEE TASLP 2008

[2] J. Kim, B. Tomasik, and D. Turnbull. *Using Artist Similarity to Propogated Semantic Information.* ISMIR 2009

[3] P.E. Rossi and R. McCulloch. *Bayesm R Package.* `http://faculty.chicagogsb.edu/peter.rossi/research/bsm.html`